STAR *fileCatalog, tagDB*
and Grand Challenge Architecture

A. Vaniachine

presenting for the Grand Challenge Collaboration

(http:/www-rnc.lbl.gov/GC/)

April 9, 2000

Joint ALICE-STAR Computing Meeting

Outline

- **GCA Overview**
- **STAR Interface:**
  - ***fileCatalog***
  - ***tagDB***
  - ***StChallenger***
- **Current Status**
- **Conclusion**

GCA: Grand Challenge Architecture

- *An order-optimized prefetch architecture for data retrieval from multilevel storage in a multiuser environment*
- Queries select events and specific event components based upon tag attribute ranges
  - query estimates are provided prior to execution
  - collections as queries are also supported
- Because event components are distributed over several files, processing an event requires delivery of a “bundle” of files
- Events are delivered in an order that takes advantage of what is already on disk, and multiuser policy-based prefetching of further data from tertiary storage
- GCA intercomponent communication is CORBA-based, but physicists are shielded from this layer

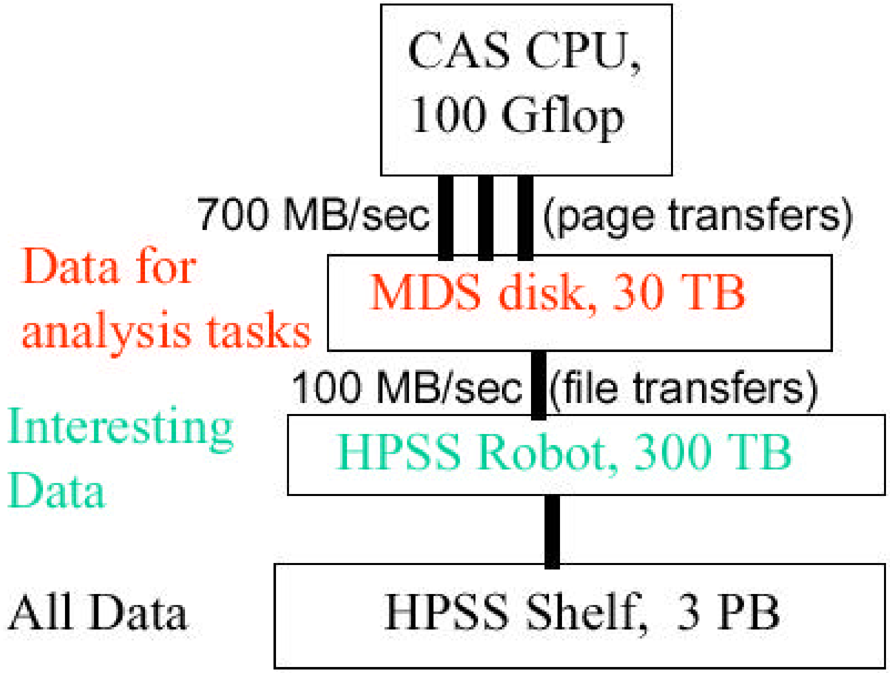
Participants

- **NERSC/Berkeley Lab**
  - L. Bernardo, A. Mueller, H. Nordberg, A.Shoshani, A. Sim, J. Wu
- **Argonne**
  - D. Malon, E. May, G. Pandola
- **Brookhaven Lab**
  - B. Gibbard, S. Johnson, J. Porter, T.Wenaus
- **Nuclear Science/Berkeley Lab**
  - D. Olson, A. Vaniachine, J. Yang, D.Zimmerman

Problem

- There are several
  - – Not all data fits on disk ($$)
    - Part of 1 year's DST's fit on disk
      - – What about last year, 2 year's ago?
      - – What about hits, raw?
  - – Available disk bandwidth means data read into memory must be efficiently used ($$)
    - don't read unused portions of the event
    - Don't read events you don't need
  - – Available tape bandwidth means files read from tape must be shared by many users, files should not contain unused bytes ($$$$)
  - – Facility resources are sufficient only if used efficiently
    - Should operate steady-state (nearly) fully loaded

Bottleneks



CAS CPU, 100 Gflop

700 MB/sec (page transfers)

Data for analysis tasks

MDS disk, 30 TB

100 MB/sec (file transfers)

Interesting Data

HPSS Robot, 300 TB

All Data

HPSS Shelf, 3 PB

Bulk bandwidth numbers meet estimated requirements assuming 100% efficiency.

How to achieve bulk bandwidth?

What fraction of data transfered is useful to programs?!!!

Keep recently accessed data on disk, but manage it so unused data does not waste space.

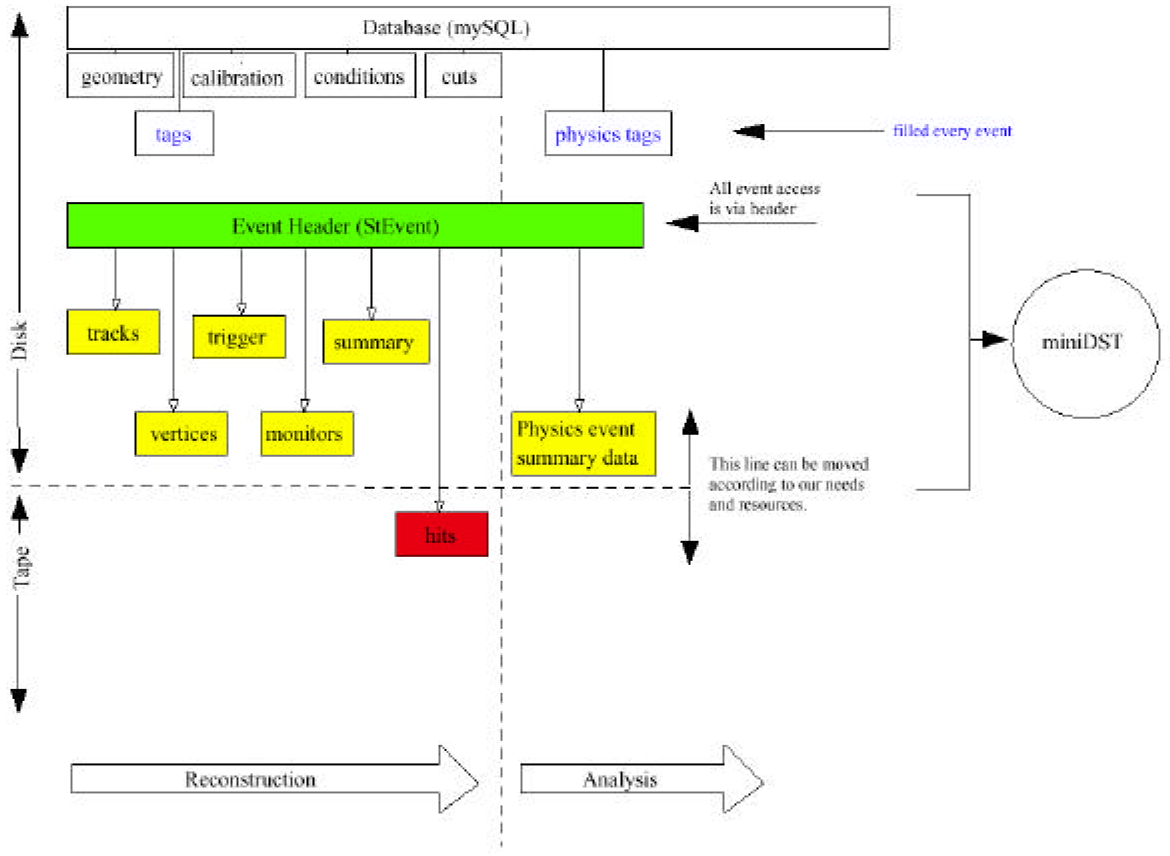
Try to arrange that 90% of file access is to disk and only 10% are retrieved from tape.

Solution Components

- Split event into components across different files so that most bytes read are used
  - Raw, tracks, hits, tags, summary, trigger, …
- Optimize file size so tape bandwidth is not wasted
  - 1GB files, → means different # of events in each file
- Coordinate file usage so tape access is shared
  - Users select all files at once
  - System optimizes retrieval and order of processing
- Use disk space & bandwidth efficiently
  - Operate disk as cache in front of tape

Grand Challenge

STAR Event Model



🟩 Parts of the data which have to be on the miniDST

🟨 Data which might or might not end up on the miniDST

🟥 Data which is unlikely to be stored on a miniDST

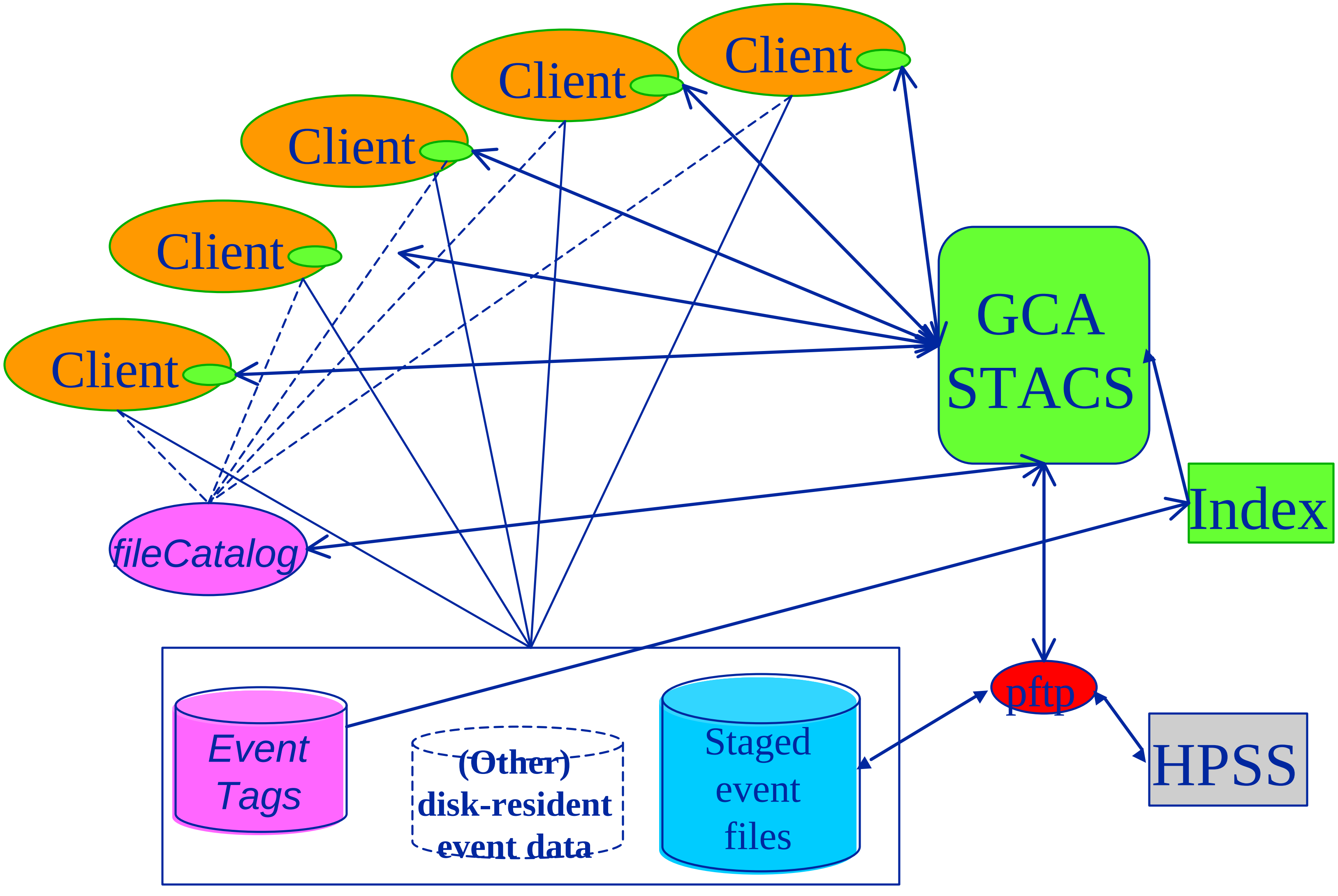
T. Ullrich, Jan. 2000

Analysis of Events

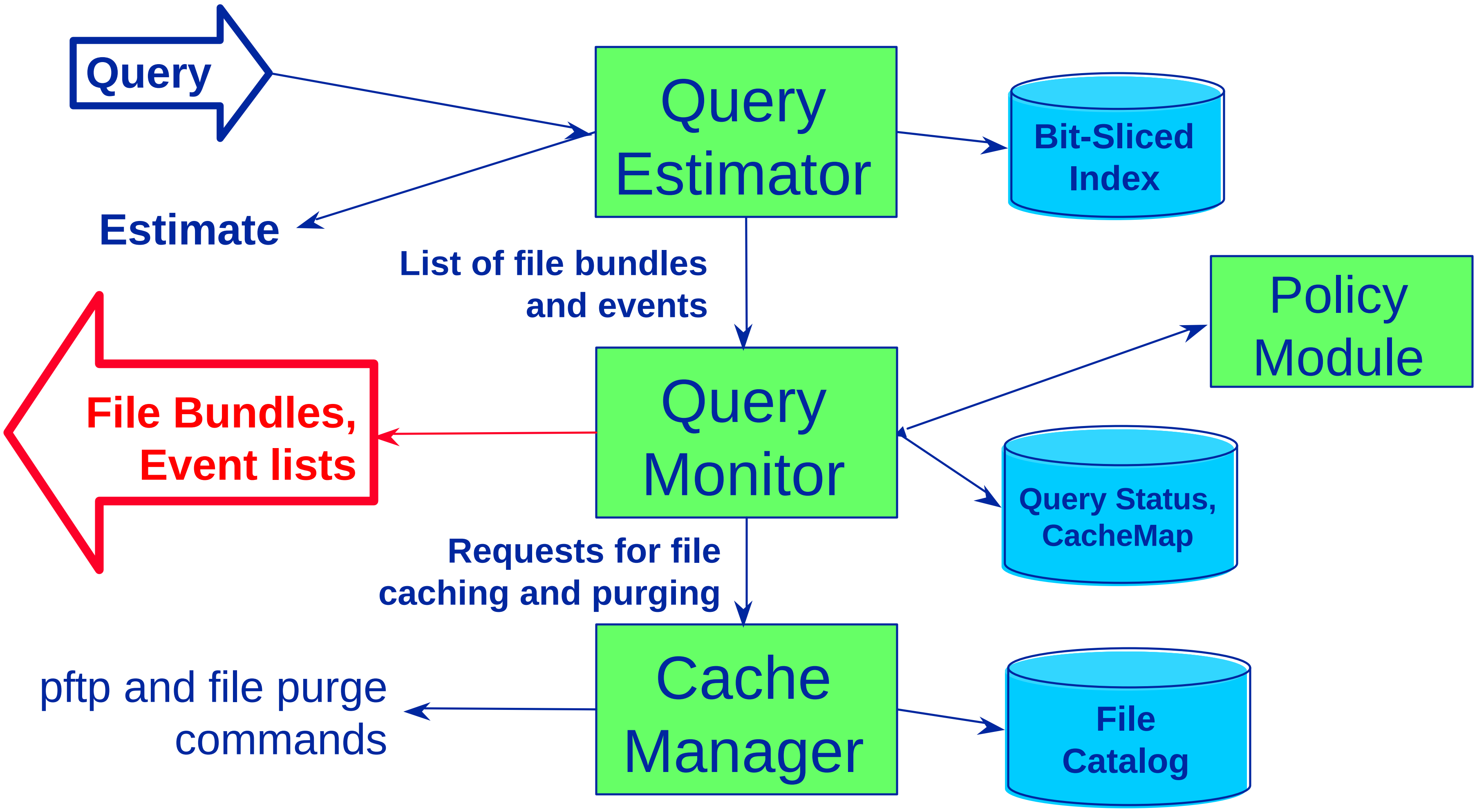
- 1M events = 100GB – 1TB
  - 100 – 1000 files (or more if not optimized)
- Need to coordinate event associations across files
- Probably have filtered some % of events
  - Suppose 25% failed cuts after trigger selection
    - Increase speed by not reading these 25%
- Run several batch jobs for same analysis in parallel to increase throughput
- Start processing with files already on disk without waiting for staging from HPSS

Grand
Challenge

GCA System Overview



STACS: STorage Access Coordination System



Query

Query Estimator

Bit-Sliced Index

Estimate

List of file bundles and events

Policy Module

File Bundles, Event lists

Query Monitor

Query Status, CacheMap

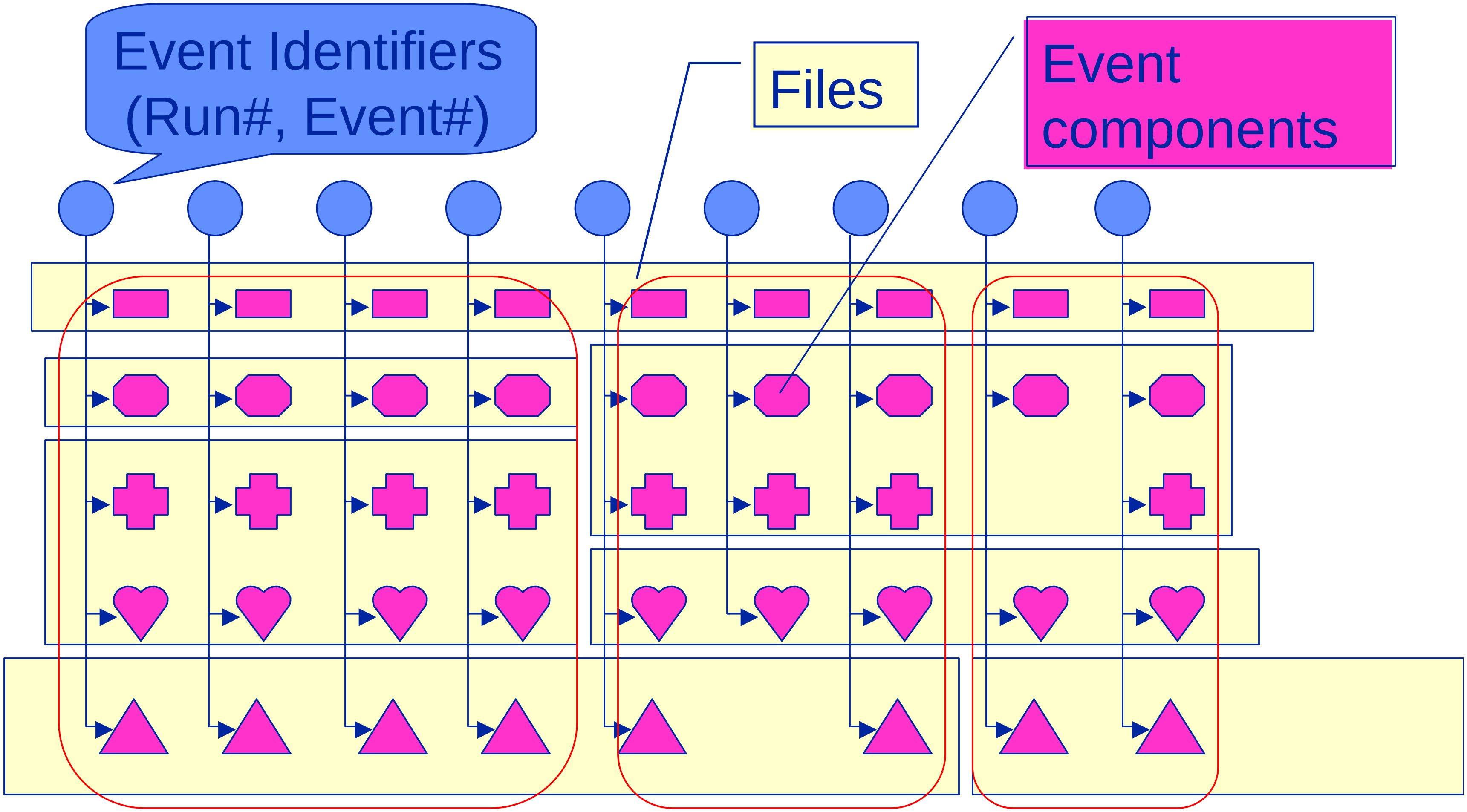
Requests for file caching and purging

pftp and file purge commands

Cache Manager

File Catalog

Organization of Events in Files



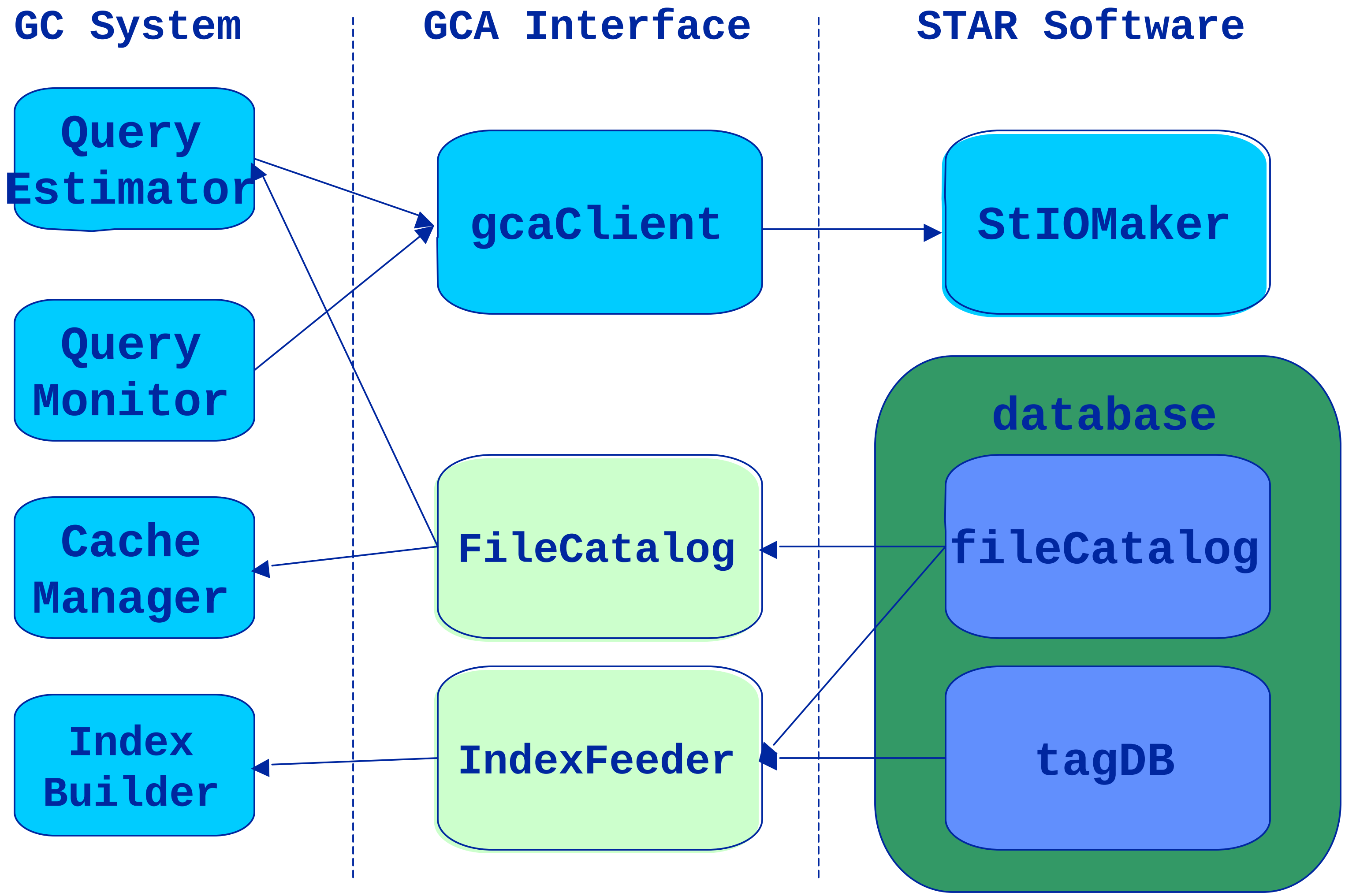
File bundle 1 File bundle 2 File bundle 3

The Details

- Range-query language, or query by event list
  - “NLa>700 && run=101007”,
  - {e1,r27012;e3,r27014;e7;r27017 …}
  - Select components: dst, geant, …
- Query estimation
  - # events, # files, # files on disk, how long, …
  - Avoid executing incorrect queries
- Order optimization
  - Order of events you get maximizes file sharing and minimizes reads from HPSS
- Policies
  - # of pre-fetch, # queries/user, # active pftp connections, …
  - Tune behavior & performance
- Parallel processing
  - Submitting same query token in several jobs will cause each job to process part of that query

Grand Challenge

Interfacing GCA to STAR



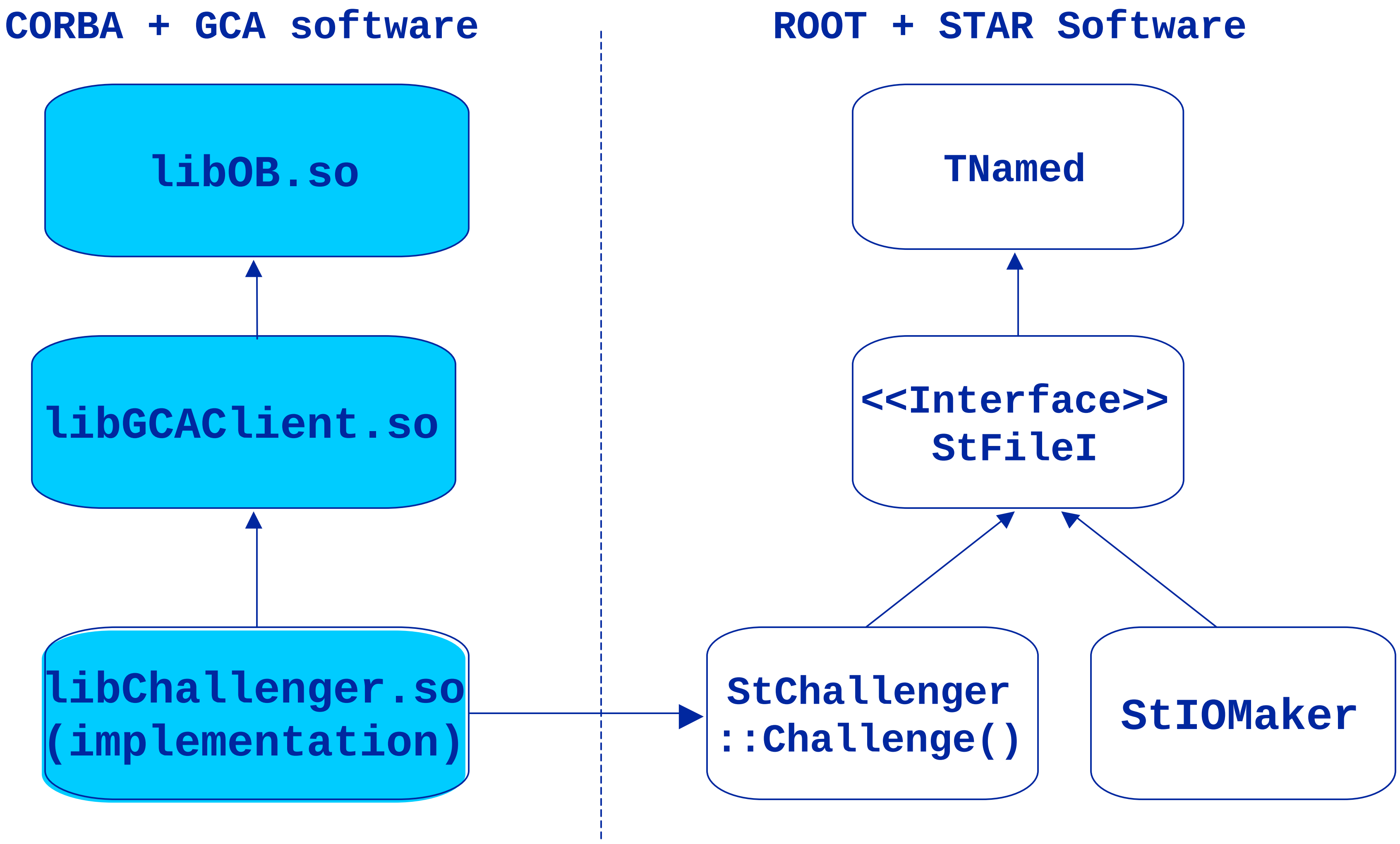
fileCatalog, tagDB and GCA

Limiting Dependencies

STAR-specific & GCA-dependent

- **IndexFeeder server**
  - **IndexFeeder read the “tag database” so that GCA “index builder” can create index**
- **FileCatalog server**
  - **FileCatalog queries the “file catalog” database of the experiment to translate fileID to HPSS & disk path**
- **gcaClient interface**
  - **Experiment sends queries and get back filenames through the gcaClient library calls**

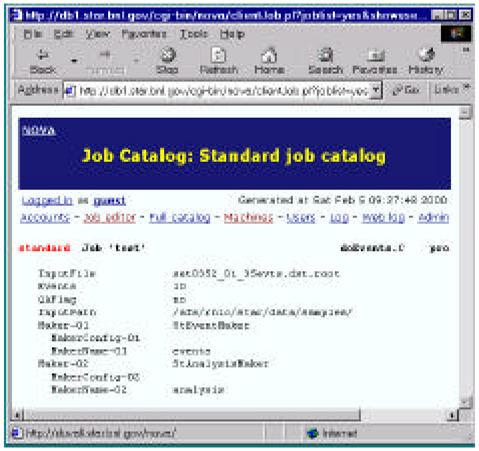
Eliminating Dependencies



STAR *fileCatalog*

- Database of information for files in experiment. File information is added to DB as files are created.
- Source of File information
  - for the experiment
  - for the GCA components (Index, gcaClient,...)

Cataloguing Analysis Workflow



http://db1.star.bnl.gov/cgi-bin/nova/clientJob.pl?joblist=yes&showon...

Address: http://db1.star.bnl.gov/cgi-bin/nova/clientJob.pl?joblist=yes

NOVA

Job Catalog: Standard job catalog

Logged in as guest — Generated at Sat Feb 5 09:27:48 2000

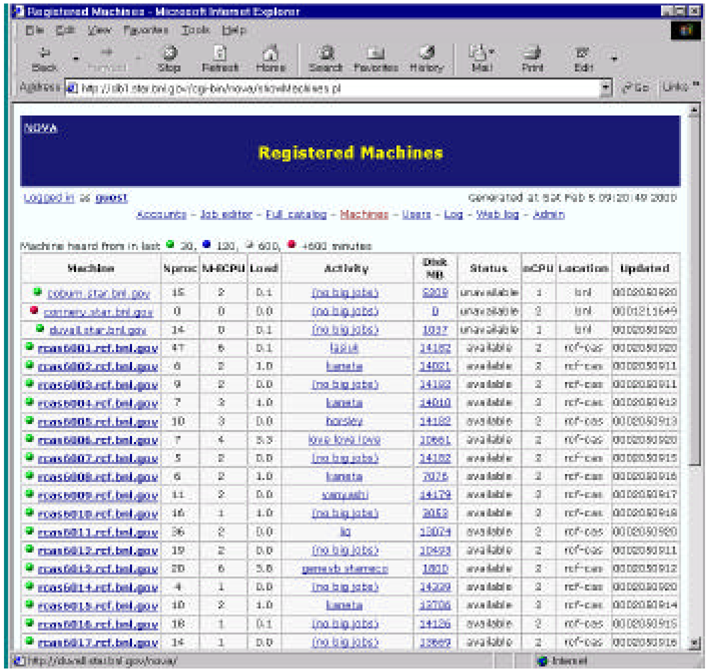
Accounts - Job editor - Full catalog - Machines - Users - Log - Web log - Admin

```
standard  Job 'test'                                  doEvents.C    pro

   InputFile           set0352_01_35evts.dst.root
   Events              10
   QAFlag              no
   InputPath           /afs/rhic/star/data/samples/
   Maker-01            StEventMaker
     MakerConfig-01
     MakerName-01      events
   Maker-02            StAnalysisMaker
     MakerConfig-02
     MakerName-02      analysis
```

http://duvall.star.bnl.gov/nova/ — Internet

Job configuration manager



Registered Machines - Microsoft Internet Explorer

Address: http://db1.star.bnl.gov/cgi-bin/nova/showMachines.pl

NOVA

Registered Machines

Logged in as guest — Generated at Sat Feb 5 09:20:49 2000

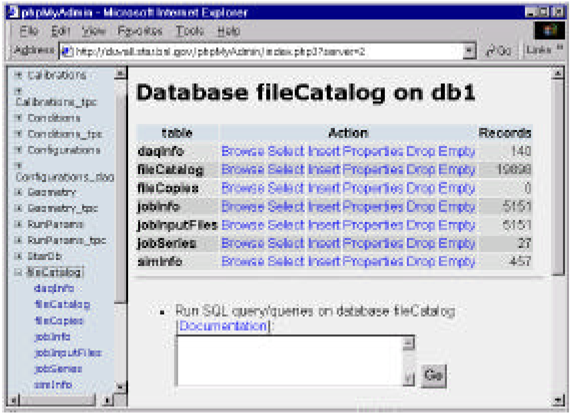
Accounts - Job editor - Full catalog - Machines - Users - Log - Web log - Admin

Machine heard from in last ● 30, ● 120, ● 600, ● +600 minutes

Machine	Nproc	N-RCPU	Load	Activity	Disk MB	Status	nCPU	Location	Updated
● coburn.star.bnl.gov	15	2	0.1	(no big jobs)	5309	unavailable	1	bnl	0002050920
● connery.star.bnl.gov	0	0	0.0	(no big jobs)	0	unavailable	2	bnl	0001211649
● duvall.star.bnl.gov	14	0	0.1	(no big jobs)	1037	unavailable	1	bnl	0002050920
● rcas6001.rcf.bnl.gov	47	6	0.1	lasiuk	14182	available	2	rcf-cas	0002050920
● rcas6002.rcf.bnl.gov	6	2	1.0	kaneta	14021	available	2	rcf-cas	0002050911
● rcas6003.rcf.bnl.gov	9	2	0.0	(no big jobs)	14182	available	2	rcf-cas	0002050911
● rcas6004.rcf.bnl.gov	7	3	1.0	kaneta	14010	available	3	rcf-cas	0002050912
● rcas6005.rcf.bnl.gov	10	3	0.0	horsley	14182	available	2	rcf-cas	0002050913
● rcas6006.rcf.bnl.gov	7	4	3.3	love love love	10661	available	2	rcf-cas	0002050920
● rcas6007.rcf.bnl.gov	5	2	0.0	(no big jobs)	14182	available	2	rcf-cas	0002050915
● rcas6008.rcf.bnl.gov	6	2	1.0	kaneta	7026	available	2	rcf-cas	0002050916
● rcas6009.rcf.bnl.gov	11	2	0.0	vanyashi	14179	available	3	rcf-cas	0002050917
● rcas6010.rcf.bnl.gov	16	1	1.0	(no big jobs)	2053	available	2	rcf-cas	0002050918
● rcas6011.rcf.bnl.gov	36	2	0.0	lq	13024	available	2	rcf-cas	0002050920
● rcas6012.rcf.bnl.gov	19	2	0.0	(no big jobs)	10493	available	2	rcf-cas	0002050911
● rcas6013.rcf.bnl.gov	20	6	3.8	genevb starreco	1800	available	2	rcf-cas	0002050912
● rcas6014.rcf.bnl.gov	4	1	0.0	(no big jobs)	14229	available	2	rcf-cas	0002050920
● rcas6015.rcf.bnl.gov	10	2	1.0	kaneta	12706	available	2	rcf-cas	0002050914
● rcas6016.rcf.bnl.gov	18	1	0.1	(no big jobs)	14126	available	2	rcf-cas	0002050915
● rcas6017.rcf.bnl.gov	14	1	0.0	(no big jobs)	13669	available	2	rcf-cas	0002050916

http://duvall.star.bnl.gov/nova/ — Internet

Job monitoring system



phpMyAdmin - Microsoft Internet Explorer

Address: http://duvall.star.bnl.gov/phpMyAdmin/index.php3?server=2

Calibrations
Calibrations_tpc
Conditions
Conditions_tpc
Configurations
Configurations_daq
Geometry
Geometry_tpc
RunParams
RunParams_tpc
StarDb
fileCatalog
daqInfo
fileCatalog
fileCopies
jobInfo
jobInputFiles
jobSeries
simInfo

Database fileCatalog on db1

table	Action	Records
daqInfo	Browse Select Insert Properties Drop Empty	140
fileCatalog	Browse Select Insert Properties Drop Empty	19898
fileCopies	Browse Select Insert Properties Drop Empty	0
jobInfo	Browse Select Insert Properties Drop Empty	5151
jobInputFiles	Browse Select Insert Properties Drop Empty	5151
jobSeries	Browse Select Insert Properties Drop Empty	27
simInfo	Browse Select Insert Properties Drop Empty	457

- Run SQL query/queries on database fileCatalog [Documentation]:

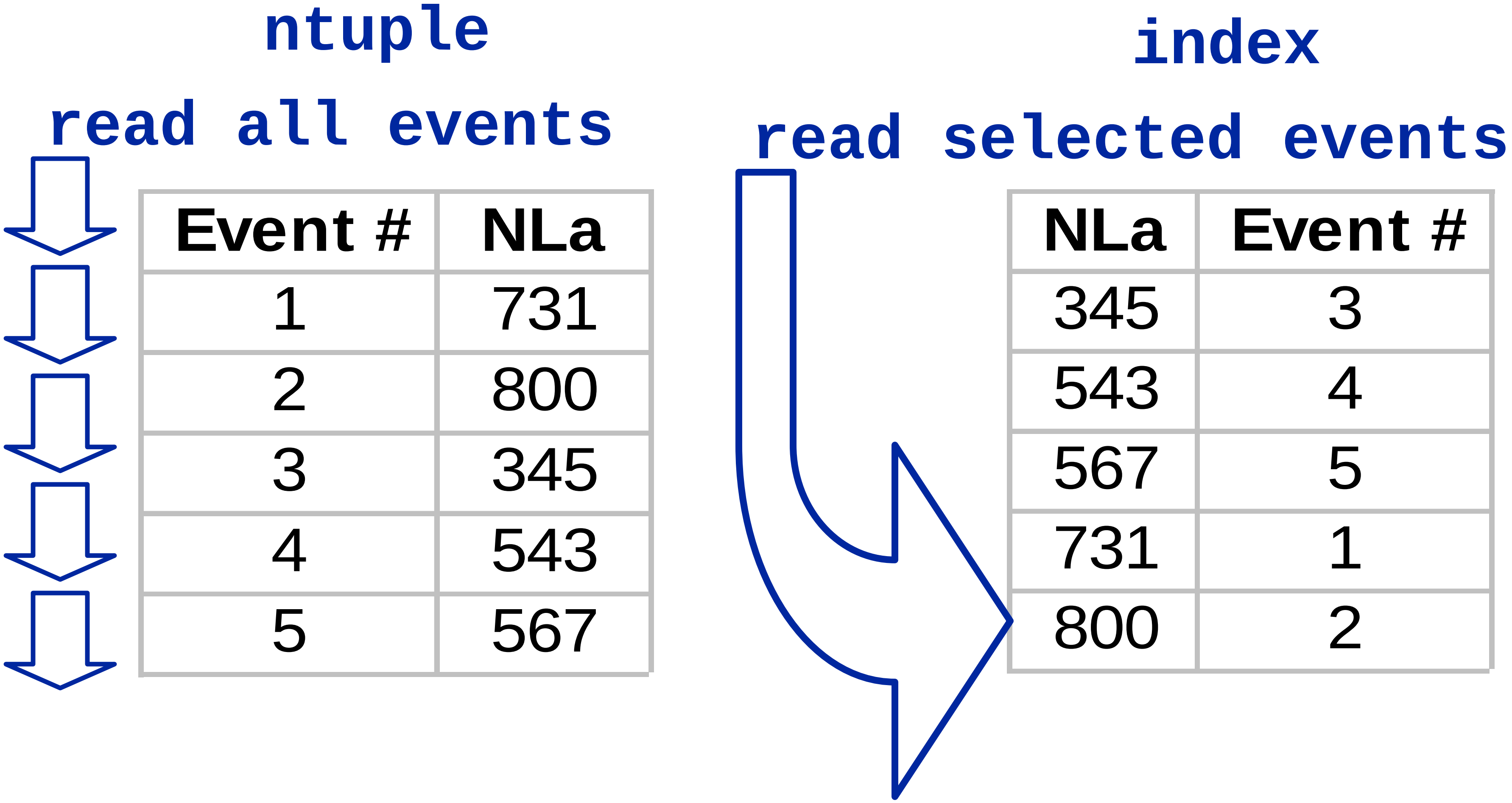
Go

fileCatalog

Transactionless Solution

- **MySQL:**
  - **no views**
    - reader has to join tables
  - **no transactions**
    - db snapshot during the update may be inconsistent
    - updates may be long (= no table locking for read)
- Experiment: few writers, hundreds of readers
- Compromise:
  - **use pre-calculated views (= extra tables)**
    - **fileCatalog data are duplicated**
  - **views update is quick (table can be locked)**
    - **clients read see consistent data at any time**

Problem: SELECT NLa>700



ntuple

read all events

Event #	NLa
1	731
2	800
3	345
4	543
5	567

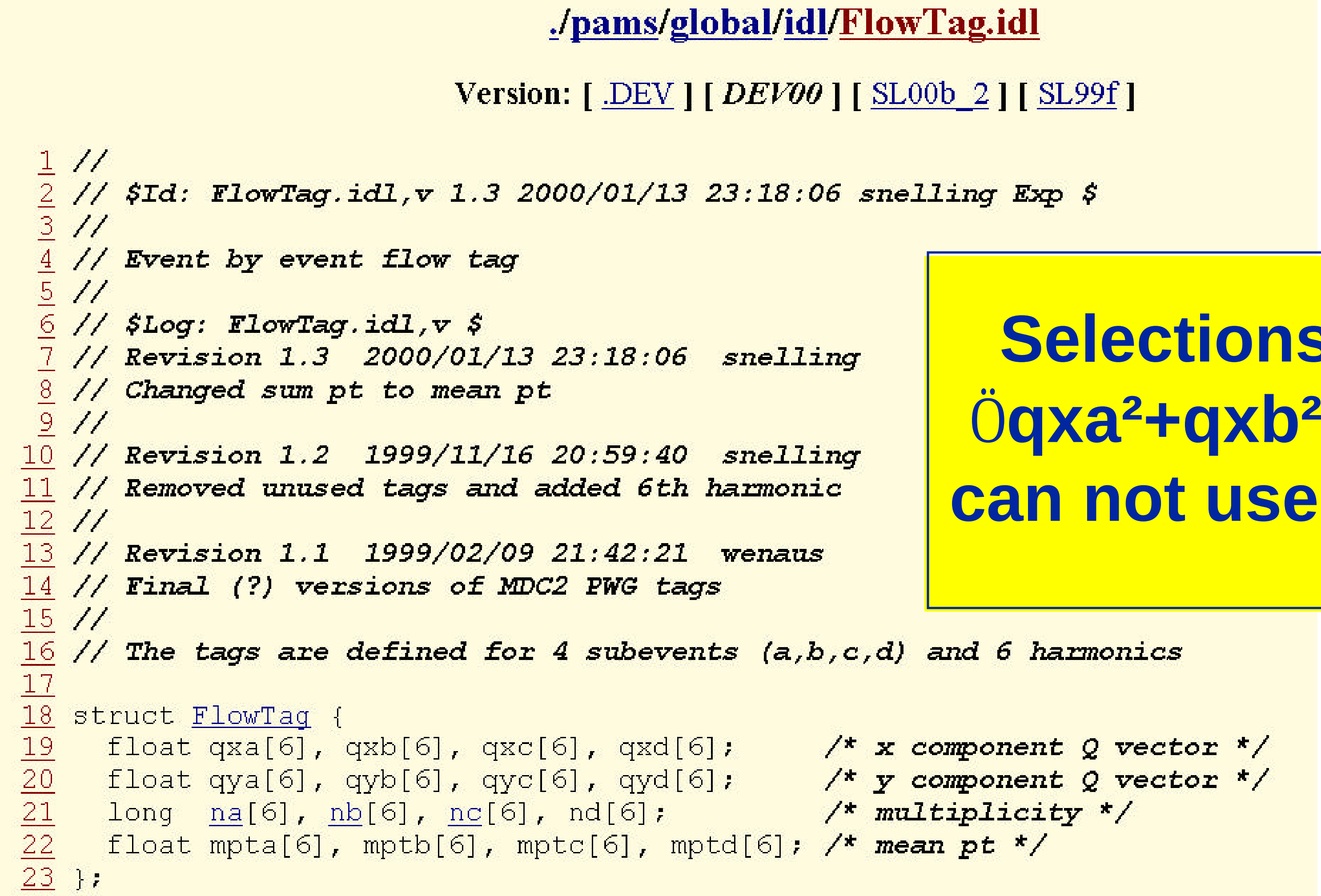
index

read selected events

NLa	Event #
345	3
543	4
567	5
731	1
800	2

Grand Challenge

STAR Tag Structure Definition



./pams/global/idl/FlowTag.idl

Version: [.DEV] [*DEV00*] [SL00b_2] [SL99f]

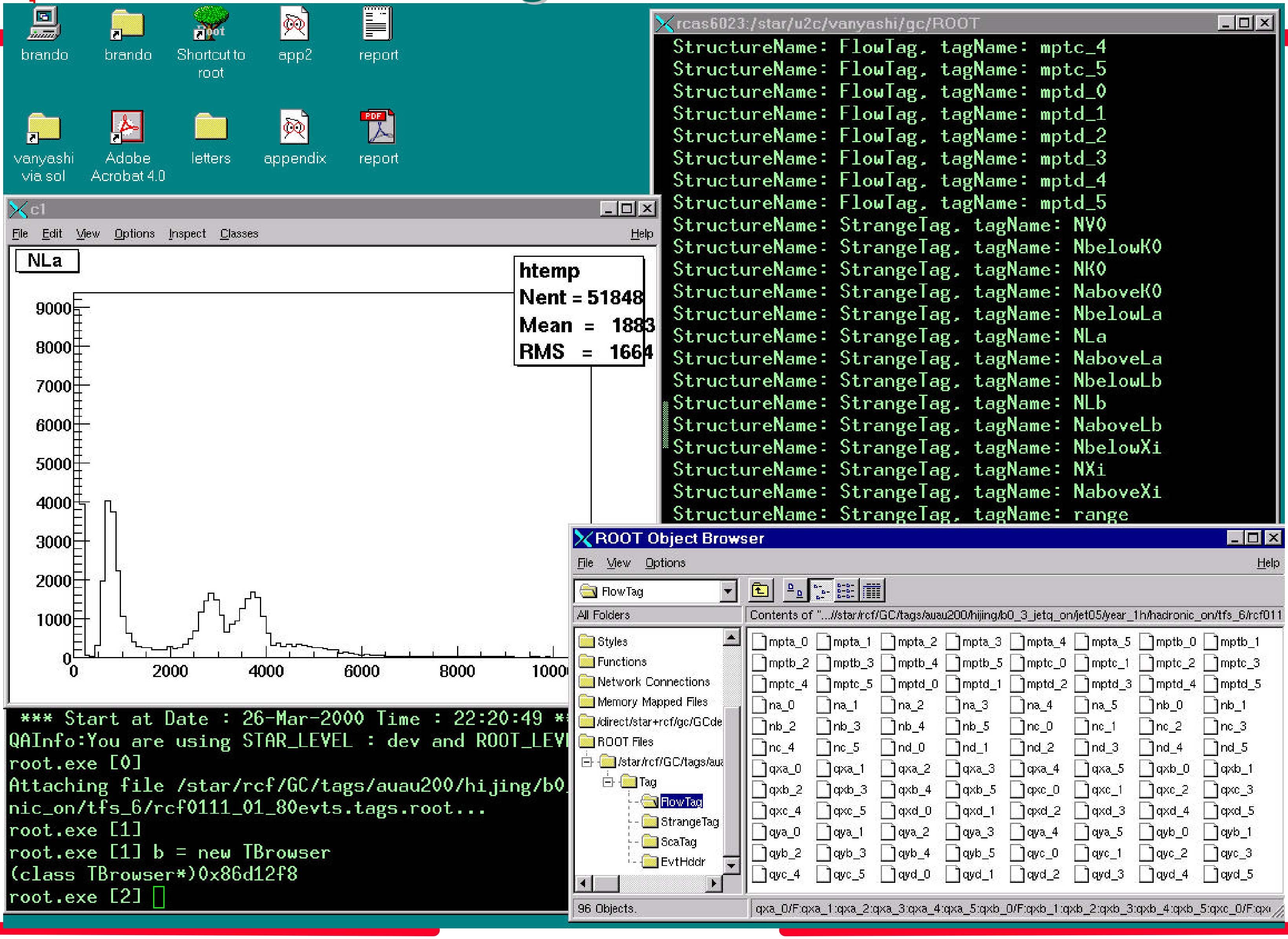
```
//
// $Id: FlowTag.idl,v 1.3 2000/01/13 23:18:06 snelling Exp $
//
// Event by event flow tag
//
// $Log: FlowTag.idl,v $
// Revision 1.3  2000/01/13 23:18:06  snelling
// Changed sum pt to mean pt
//
// Revision 1.2  1999/11/16 20:59:40  snelling
// Removed unused tags and added 6th harmonic
//
// Revision 1.1  1999/02/09 21:42:21  wenaus
// Final (?) versions of MDC2 PWG tags
//
// The tags are defined for 4 subevents (a,b,c,d) and 6 harmonics

struct FlowTag {
  float qxa[6], qxb[6], qxc[6], qxd[6];     /* x component Q vector */
  float qya[6], qyb[6], qyc[6], qyd[6];     /* y component Q vector */
  long  na[6], nb[6], nc[6], nd[6];         /* multiplicity */
  float mpta[6], mptb[6], mptc[6], mptd[6]; /* mean pt */
};
```

Selections like Ö$qxa^2+qxb^2 > 0.5$ can not use index

Grand Challenge

STAR Tag Database Access



```
StructureName: FlowTag, tagName: mptc_4
StructureName: FlowTag, tagName: mptc_5
StructureName: FlowTag, tagName: mptd_0
StructureName: FlowTag, tagName: mptd_1
StructureName: FlowTag, tagName: mptd_2
StructureName: FlowTag, tagName: mptd_3
StructureName: FlowTag, tagName: mptd_4
StructureName: FlowTag, tagName: mptd_5
StructureName: StrangeTag, tagName: NV0
StructureName: StrangeTag, tagName: NbelowK0
StructureName: StrangeTag, tagName: NK0
StructureName: StrangeTag, tagName: NaboveK0
StructureName: StrangeTag, tagName: NbelowLa
StructureName: StrangeTag, tagName: NLa
StructureName: StrangeTag, tagName: NaboveLa
StructureName: StrangeTag, tagName: NbelowLb
StructureName: StrangeTag, tagName: NLb
StructureName: StrangeTag, tagName: NaboveLb
StructureName: StrangeTag, tagName: NbelowXi
StructureName: StrangeTag, tagName: NXi
StructureName: StrangeTag, tagName: NaboveXi
StructureName: StrangeTag, tagName: range
```

```
 *** Start at Date : 26-Mar-2000 Time : 22:20:49 **
QAInfo:You are using STAR_LEVEL : dev and ROOT_LEVE
root.exe [0]
Attaching file /star/rcf/GC/tags/auau200/hijing/b0_
nic_on/tfs_6/rcf0111_01_80evts.tags.root...
root.exe [1]
root.exe [1] b = new TBrowser
(class TBrowser*)0x86d12f8
root.exe [2] 
```

Contents of ".../star/rcf/GC/tags/auau200/hijing/b0_3_jetq_on/jet05/year_1h/hadronic_on/tfs_6/rcf011

```
mpta_0  mpta_1  mpta_2  mpta_3  mpta_4  mpta_5  mptb_0  mptb_1
mptb_2  mptb_3  mptb_4  mptb_5  mptc_0  mptc_1  mptc_2  mptc_3
mptc_4  mptc_5  mptd_0  mptd_1  mptd_2  mptd_3  mptd_4  mptd_5
na_0    na_1    na_2    na_3    na_4    na_5    nb_0    nb_1
nb_2    nb_3    nb_4    nb_5    nc_0    nc_1    nc_2    nc_3
nc_4    nc_5    nd_0    nd_1    nd_2    nd_3    nd_4    nd_5
qxa_0   qxa_1   qxa_2   qxa_3   qxa_4   qxa_5   qxb_0   qxb_1
qxb_2   qxb_3   qxb_4   qxb_5   qxc_0   qxc_1   qxc_2   qxc_3
qxc_4   qxc_5   qxd_0   qxd_1   qxd_2   qxd_3   qxd_4   qxd_5
qya_0   qya_1   qya_2   qya_3   qya_4   qya_5   qyb_0   qyb_1
qyb_2   qyb_3   qyb_4   qyb_5   qyc_0   qyc_1   qyc_2   qyc_3
qyc_4   qyc_5   qyd_0   qyd_1   qyd_2   qyd_3   qyd_4   qyd_5
```

qxa_0/F:qxa_1:qxa_2:qxa_3:qxa_4:qxa_5:qxb_0/F:qxb_1:qxb_2:qxb_3:qxb_4:qxb_5:qxc_0/F:qx

TagDB in ROOT Files

- **Tag data are stored in ROOT TTree files**
- **Branches (3 out of 6 requested) saved in the split mode**
    - StrangeTag
    - FlowTag
    - ScaTag
- **173 physics tags [int/float] out of 500 requested**
- **Disk resident tag files name+path are stored in the MySQL *fileCatalog* database**
- **Files are selected from database in the ROOT CINT macro through the ROOT-MySQL interface and are chained for further selections by user**

MDC3 Index

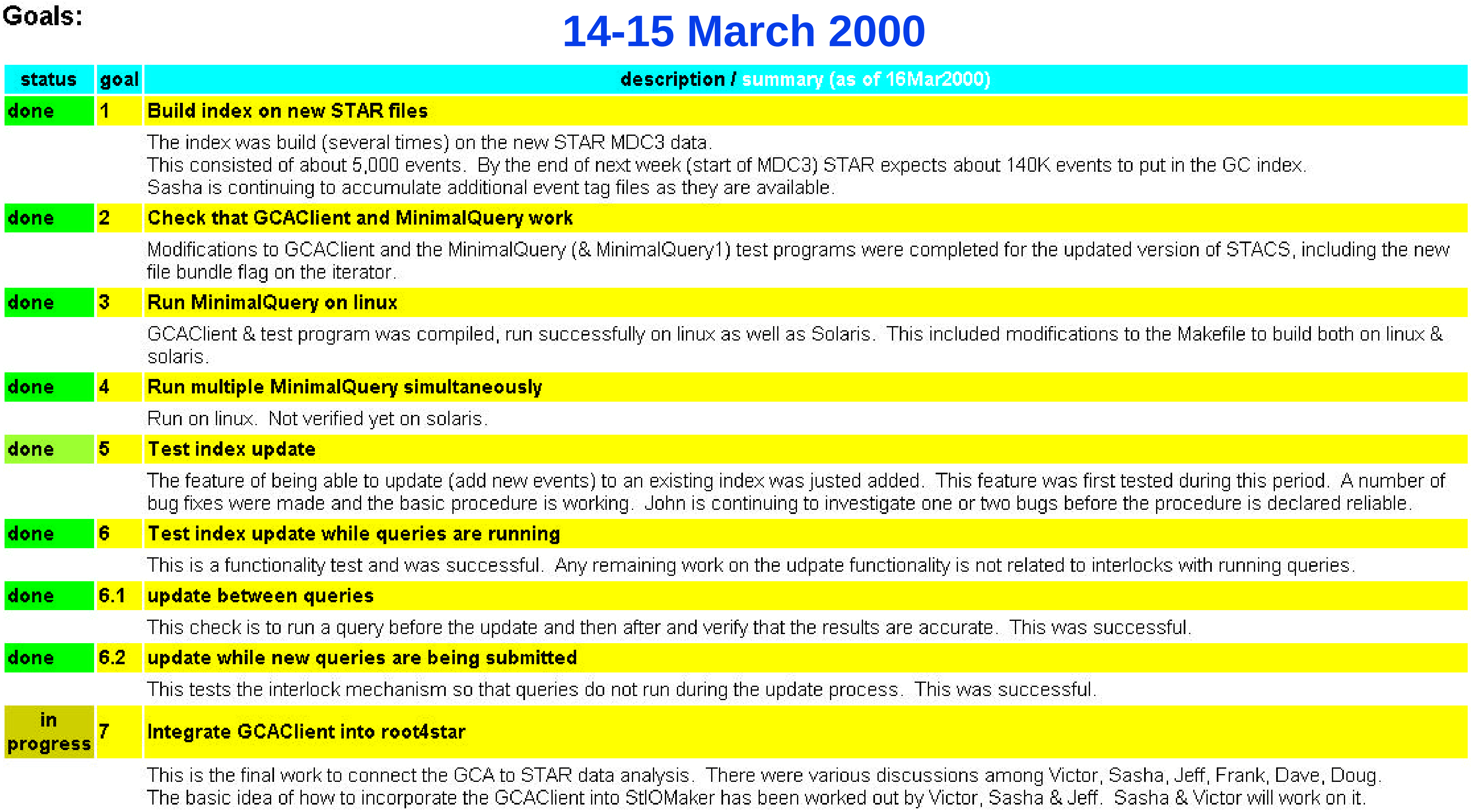
- 6 event components:
  - fzd
  - geant
  - dst
  - tags
  - runco
  - hist
- 179 physics tags:
  - StrangeTag
  - FlowTag
  - ScaTag
- 120K events
- 8K files

Grand Challenge

GCA MDC3 Integration Workshop

http://www-rnc.lbl.gov/GC/meetings/14mar00/default.htm

14-15 March 2000



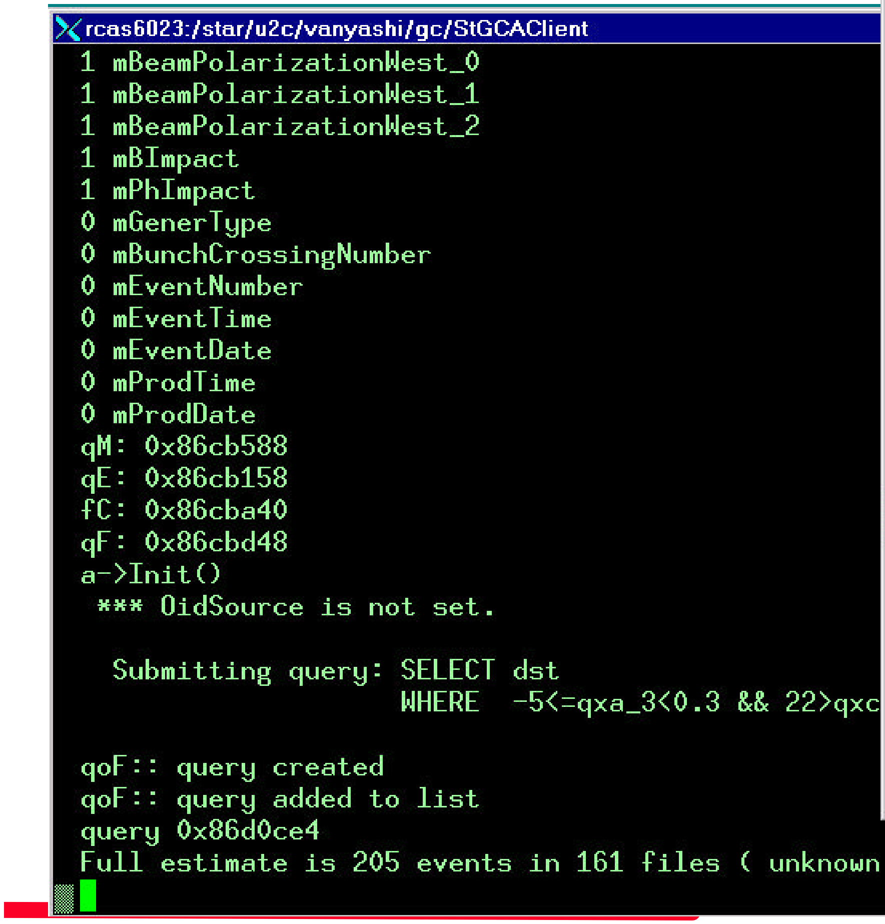
Goals:

status	goal	description / summary (as of 16Mar2000)
done	1	**Build index on new STAR files**
		The index was build (several times) on the new STAR MDC3 data. This consisted of about 5,000 events. By the end of next week (start of MDC3) STAR expects about 140K events to put in the GC index. Sasha is continuing to accumulate additional event tag files as they are available.
done	2	**Check that GCAClient and MinimalQuery work**
		Modifications to GCAClient and the MinimalQuery (& MinimalQuery1) test programs were completed for the updated version of STACS, including the new file bundle flag on the iterator.
done	3	**Run MinimalQuery on linux**
		GCAClient & test program was compiled, run successfully on linux as well as Solaris. This included modifications to the Makefile to build both on linux & solaris.
done	4	**Run multiple MinimalQuery simultaneously**
		Run on linux. Not verified yet on solaris.
done	5	**Test index update**
		The feature of being able to update (add new events) to an existing index was justed added. This feature was first tested during this period. A number of bug fixes were made and the basic procedure is working. John is continuing to investigate one or two bugs before the procedure is declared reliable.
done	6	**Test index update while queries are running**
		This is a functionality test and was successful. Any remaining work on the udpate functionality is not related to interlocks with running queries.
done	6.1	**update between queries**
		This check is to run a query before the update and then after and verify that the results are accurate. This was successful.
done	6.2	**update while new queries are being submitted**
		This tests the interlock mechanism so that queries do not run during the update process. This was successful.
in progress	7	**Integrate GCAClient into root4star**
		This is the final work to connect the GCA to STAR data analysis. There were various discussions among Victor, Sasha, Jeff, Frank, Dave, Doug. The basic idea of how to incorporate the GCAClient into StIOMaker has been worked out by Victor, Sasha & Jeff. Sasha & Victor will work on it.

Grand Challenge

User Query

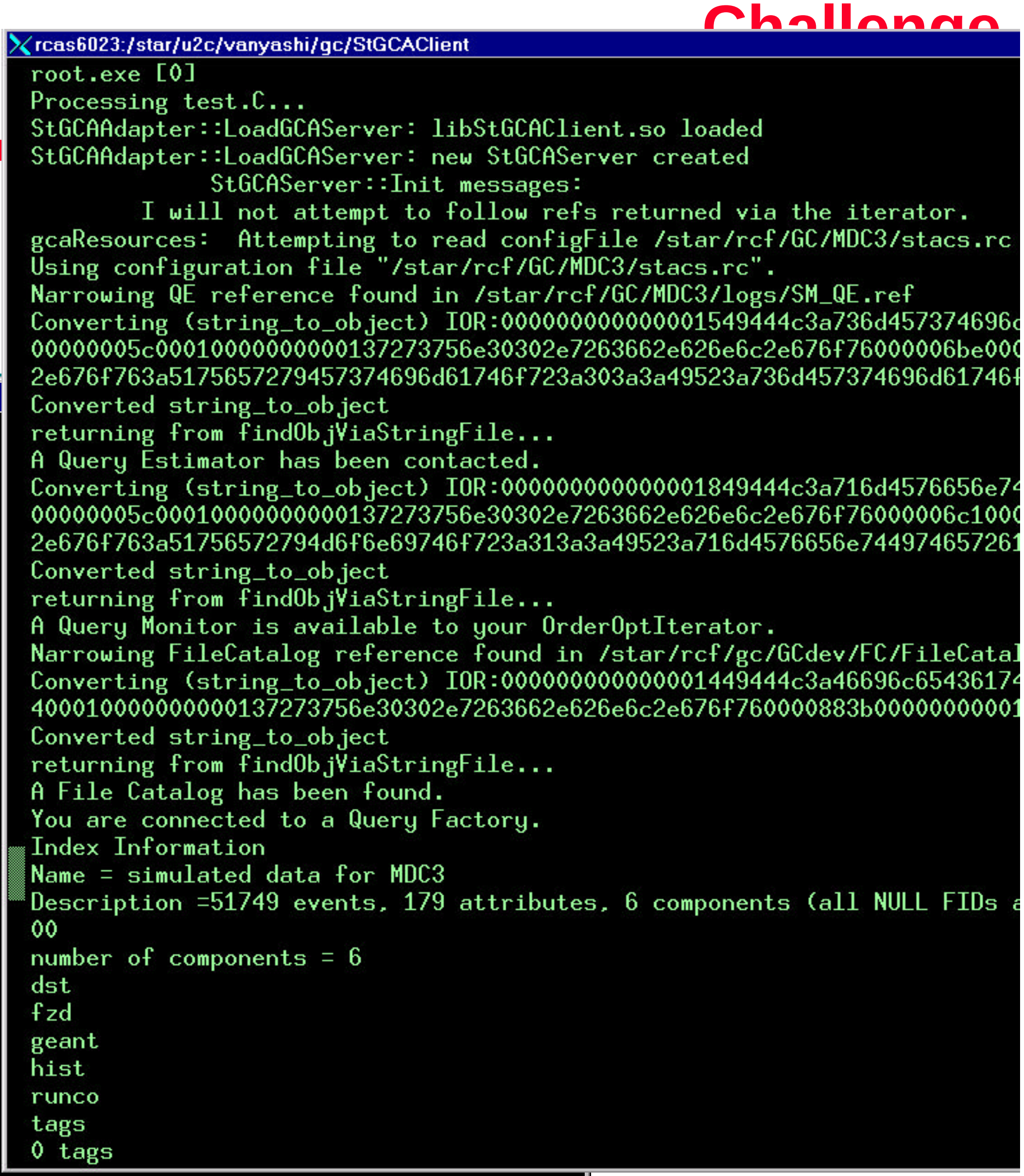
ROOT Session:



```
1 mBeamPolarizationWest_0
1 mBeamPolarizationWest_1
1 mBeamPolarizationWest_2
1 mBImpact
1 mPhImpact
0 mGenerType
0 mBunchCrossingNumber
0 mEventNumber
0 mEventTime
0 mEventDate
0 mProdTime
0 mProdDate
qM: 0x86cb588
qE: 0x86cb158
fC: 0x86cba40
qF: 0x86cbd48
a->Init()
 *** OidSource is not set.

  Submitting query: SELECT dst
                    WHERE  -5<=qxa_3<0.3 && 22>qxc

qoF:: query created
qoF:: query added to list
query 0x86d0ce4
Full estimate is 205 events in 161 files ( unknown MBs).
```



```
root.exe [0]
Processing test.C...
StGCAAdapter::LoadGCAServer: libStGCAClient.so loaded
StGCAAdapter::LoadGCAServer: new StGCAServer created
             StGCAServer::Init messages:
        I will not attempt to follow refs returned via the iterator.
gcaResources:  Attempting to read configFile /star/rcf/GC/MDC3/stacs.rc
Using configuration file "/star/rcf/GC/MDC3/stacs.rc".
Narrowing QE reference found in /star/rcf/GC/MDC3/logs/SM_QE.ref
Converting (string_to_object) IOR:000000000000001549444c3a736d457374696c
00000005c00010000000000137273756e30302e7263662e626e6c2e676f76000006be000
2e676f763a5175657279457374696d61746f723a303a3a49523a736d457374696d61746f
Converted string_to_object
returning from findObjViaStringFile...
A Query Estimator has been contacted.
Converting (string_to_object) IOR:000000000000001849444c3a716d4576656e74
00000005c00010000000000137273756e30302e7263662e626e6c2e676f76000006c1000
2e676f763a51756572794d6f6e69746f723a313a3a49523a716d4576656e744497465726 1
Converted string_to_object
returning from findObjViaStringFile...
A Query Monitor is available to your OrderOptIterator.
Narrowing FileCatalog reference found in /star/rcf/gc/GCdev/FC/FileCatal
Converting (string_to_object) IOR:000000000000001449444c3a46696c65436174
40001000000000137273756e30302e7263662e626e6c2e676f760000883b00000000001
Converted string_to_object
returning from findObjViaStringFile...
A File Catalog has been found.
You are connected to a Query Factory.
Index Information
Name = simulated data for MDC3
Description =51749 events, 179 attributes, 6 components (all NULL FIDs a
00
number of components = 6
dst
fzd
geant
hist
runco
tags
0 tags
```

Conclusion

- GCA developed a system for optimized access to multi-component event data files stored in HPSS
- General CORBA interfaces are defined for interfacing with the experiment
- A client component encapsulates interaction with the servers and provides an ODMG-style iterator
- Has been tested up to 10M events, 7 event components, 250 concurrent queries
- Has been integrated with the STAR experiment ROOT-based I/O analysis system